DOCUMENT RESUME

ED 339 794 CE 051 066

AUTHOR	Green, Bert F., Jr., Ed.; Wing, Hilda, Ed.
TITLE	Analysis of Job Performance Measurement Data. Report of a Workshop.
INSTITUTION	National Academy of Sciences - National Research Council, Washington, DC. Commission on Behavioral and Social Sciences and Education.
SPONS AGENCY	Office of Naval Research, Arlington, Va.; Office of the Assistant Secretary of Defense for Force Management and Personnel (DOD), Washington, DC.
PUB DATE	88
CONTRACT	N00014-83-C-0448
NOTE	60p.
AVAILABLE FROM	National Academy Press, 2101 Constitution Avenue, N.W., Washington, DC 20418 (free).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Adult Education; Armed Forces; Career Education; *Criterion Referenced Tests; *Data Analysis; Data Collection; *Evaluation Criteria; *Job Performance; *Measurement Techniques; *Military Service
IDENTIFIERS	*Army

ABSTRACT

This report describes a workshop at which Army researchers presented some results from the first phase of a two-phase Joint-Service Project. (The objective of this phase was to determine if technically adequate criterion measures can be developed that are representative of job performance.) Part I of the report presents the preliminary results of the research of the Army's Project A, a selection and classification project. It describes the data collection and the Army researchers' analysis and preliminary conclusions based on those data and concludes with the Committee on the Performance of Military Personnel's reaction to the Army's efforts, as they pertain to the Joint-Service Job Performance Measurement/Enlistment Standards Project. Part II consists of a series of recommendations to the Job Performance Measurement Working Group for a core set of analyses of the job performance data collected by the Services over the past several years. The recommendations fall into two broad categories. The first and larger set deals with scientific evaluation of the performance data, including recommendations on scoring and scaling, handling missing data, reliability analysis, and exploring score relationships. The concluding set of recommendations concerns analyses specifically designed to labor force policy decisions. (YLB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

C IOM

ED339794

Analysis of Job Performance Measurement Data

Report of a Workshop

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. M. Burns

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CE051066

BEST COPY AVAILABLE

CE

Analysis of Job Performance Measurement Data

Report of a Workshop

Bert F. Green, Jr., and Hilda Wing, *editors*

Committee on the Performance of Military Personnel
Commission on Behavioral and Social Sciences and Education
National Research Council

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1988

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Samuel O. Thier is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

The work of the Committee on the Performance of Military Personnel is sponsored by the Office of the Assistant Secretary of Defense (Force Management and Personnel) and funded under Office of Naval Research Contract N00014-83-C-0448.

Available from:

Committee on the Performance of Military Personnel
National Research Council
2101 Constitution Avenue N.W.
Washington, D.C. 20418

Printed in the United States of America

COMMITTEE ON THE PERFORMANCE OF MILITARY PERSONNEL

BERT F. GREEN, JR. (Chair), Department of Psychology, Johns Hopkins University
JERALD G. BACHMAN, Institute for Social Research, University of Michigan
V. JON BENTZ, Elmhurst, Ill.
LLOYD BOND, Learning Research and Development Center, University of Pittsburgh
RICHARD V.L. COOPER, Coopers & Lybrand, Inc., Washington, D.C.
RICHARD DANZIG, Latham & Watkins, Washington, D.C.
FRANK J. LANDY, Department of Psychology, Pennsylvania State University
ROBERT L. LINN, School of Education, University of Colorado
JOHN W. ROBERTS, San Antonio, Tex.
DONALD B. RUBIN, Department of Statistics, Harvard University
MADY W. SEGAL, Department of Sociology, University of Maryland, College Park
RICHARD J. SHAVELSON, Graduate School of Education, University of California, Santa Barbara
H.P. VAN COTT, Committee on Human Factors, National Research Council

ALEXANDRA K. WIGDOR, Study Director
HILDA WING, Research Associate
DIANE L. GOLDMAN, Administrative Secretary

Contents

INTRODUCTION 1

PART I THE ARMY'S PROJECT A RESEARCH 5

The Army's Data Collection, 7
The Army's Analysis, 23
The Army's Conclusions, 31
Committee Discussion, 31

PART II PROPOSALS FOR COMMON DATA ANALYSIS 37

Recommendations: Analyses Necessary for the Scientific Evaluation of the Joint-Service Project, 39
Recommendations: Analyses Useful for Informing Manpower Policy Decisions, 48

DEFINITIONS 49

REFERENCES 51

ATTENDANCE 53

Introduction

In 1980 it was determined that the major military entrance test, the Armed Services Vocational Aptitude Battery (ASVAB), had been misnormed, with the result that scores in the lower ranges were artificially inflated. Approximately 250,000 young men and women had entered military service between 1976 and 1980 who would have been unable otherwise to meet enlistment standards. As a consequence of the misnorming, policy makers in both Congress and the Department of Defense (DoD) became very interested in establishing the relationship of the ASVAB to actual job performance.

In response, each of the four military Services began to address the complexities of evaluating job performance. DoD supported these efforts and, at the direction of Congress, combined them into one major research program, the Joint-Service Job Performance Measurement/Enlistment Standards Project. Under the chairmanship of DoD, technical and policy representatives of the four Services formed the Joint-Service Job Performance Measurement Working Group to coordinate the research efforts of the four Services. The Committee on the Performance of Military Personnel was established within the National Research Council to provide independent, scientific assessment of the Joint-Service Project. Since 1983, the committee and the Job Performance Measurement Working Group have had a series of meetings, large and small, to explore research issues of mutual interest.

The Joint-Service Project includes two phases. The first is to determine if technically adequate criterion measures can be devel-

oped that are representative of job performance. The second is to determine how well the current enlistment procedures, including the ASVAB, predict these approximations and to develop methodologies that link the job performance measures to enlistment standards (Committee on the Performance of Military Personnel, 1984). During the past few years, the Services and DoD have received increasingly more pressure in congressional hearings to provide a much stronger and more explicitly defined data base, incorporating these two phases, to support requests for force quality requirements.

In 1983 the committee organized a workshop with members of the Job Performance Measurement Working Group as well as other representatives of the four Services. That workshop focused on the careful planning and coordination required for successful completion of the Joint-Service Project (Committee on the Performance of Military Personnel, 1984). In an interim report (Wigdor and Green, 1986), the committee concluded that the Joint-Service Project had successfully met its first objective: the development of technically adequate criterion measures that are representative of job performance.

This report describes a more recent workshop, focused on data analysis, that was held December 12-13, 1986, at the Brookshire Hotel in Baltimore, Maryland. It is divided into two parts: Part I presents the preliminary results of the Army's Project A research. It describes the data collection, the Army researchers' analysis and preliminary conclusions based on those data, and concludes with the committee's reaction to the Army's efforts, as they pertain to the Joint-Service Job Performance Measurement/Enlistment Standards Project. Part II consists of a series of recommendations to the Job Performance Measurement Working Group for a core set of analyses of the job performance data collected by the Services over the past several years.

The purpose of the 1986 workshop was to explore strategies for analyzing the voluminous amount of performance data being collected as part of the Joint-Service Project. Because the Army's research began prior to that of the other Services, in 1981, this workshop was intended to serve as an opportunity for the Army to present the experience and insights gleaned from its selection and classification project, Project A. Of most interest was the Army's experience with the construction and collection of criterion data, the goal of phase one of the Joint-Service Project. Of secondary interest was the progress made toward the goal of phase two, validation and setting enlistment standards.

The Army's research is of much broader scope than that of the Joint-Service Project. In addition to the development of performance criterion measures (including school and training success), the Army's Project A includes the development of new predictors, the development of additional criterion measures encompassing soldier effectiveness or Army-wide performance, and the validation of the new predictors and the ASVAB against the new and existing criterion measures. The utility of different levels of performance for different military occupational specialties is also being evaluated.

Accompanying Project A, the Army has another research endeavor, Project B, which is developing a computerized allocation system for optimal matching of recruits having different patterns of abilities with occupational specialties having different patterns of requirements. The new system will incorporate the findings of Project A. Its goal is to improve the level of individual job performance, by more differentiated classification techniques, both within and across occupational specialties (Schmitz, 1987).

Project A includes two major samples of soldiers, one assessed via a cross-sectional design and the other via a longitudinal design. A third and smaller longitudinal sample is included in the cross-sectional design but was evaluated prior and in addition to the larger sample. Criterion measures are being collected during both the first and second tours of duty for both samples; that is, both samples have a longitudinal component.

By the time of this workshop, the Army had completed many analyses of data from the cross-sectional sample collected during the soldiers' first tour of duty. A total of approximately 9,500 soldiers in 19 military occupational specialties had been administered both predictor and performance measures. Soldiers in nine of these occupations also received occupation-specific testing, providing the data most pertinent to the Joint-Service Project. For each soldier there were approximately 270 data values.

The Army scientists were forthcoming in their presentation of these early findings to the committee and the Job Performance Measurement Working Group. They discussed problems of missing data, construct development, construct weighting, and research results. The other Services benefited from learning what Project A had done so far and how. Results from the early stages of data analysis provided critical information; members of the Joint-Service Project could learn from the Army's successes and from its setbacks, and the Army scientists could benefit from observations from different points of view.

Part I

The Army's Project A Research

In the course of the workshop, four members of the Project A research team discussed various aspects of the data collection and analyses. Lauress L. Wise described the treatment of missing data (Wise, McHenry, and Young, 1986). He reported that the problem of missing data can be minimized by careful attention to data collected on site, to be sure that all measures are being administered properly and that all the data for each soldier are forwarded to the site data collection center. In addition, different types of measures (e.g., ratings, performance tests) are likely to have different numbers of missing data, for good reasons. A decision must be made about how many missing data are too many for each type of measure. Finally, a datum could be absent because it was, indeed, missing, or because the soldier was not attending to the task. Procedures must be developed for selecting from these two alternatives and, if the datum is missing for legitimate reasons, for imputing a value based on the other data from the same soldier. Jeffrey J. McHenry (Wise et al., 1986) talked about the requirements and procedures for developing higher-order variables, or constructs, to represent the job performance domain appropriately in each of the 19 military occupational specialties studied, particularly the 9 for which hands-on measures were administered. Of major concern was the existence of the approximately 270 variables per soldier that had to be subdivided or grouped in some intelligent fashion prior to computerized statistical analyses. These procedures can magnify chance relationships among variables and can easily produce misleading results. Robert Sadacca

spoke about obtaining the judged relative importance of these performance constructs (Sadacca, de Vera, and DiFazio, 1986). He compared several methodologies for obtaining these value judgments from both commissioned and noncommissioned officers, for different kinds and levels of performance by soldiers in different occupational specialties. Finally, John P. Campbell, principal scientist of Project A, presented an overall summary of the preliminary research results (Campbell, 1986).

This summary of the discussion draws mainly on Campbell's overview and considers the other papers in the context of the Joint-Service Project. According to Campbell (1986), the aims of Project A are:

- The development of new instruments for personnel selection and classification;
- The validation of the current predictor battery—the ASVAB—and the validation of the new instruments for predicting several aspects of job performance;
- The development of appropriate measures of job performance;
- The assessment of the differential validity of the predictor instruments for use in making decisions about classification;
- The development of materials that would allow "what if" simulations to explore, for example, the effect of using different cutoff scores in personnel decision making.

In meeting these aims, the project researchers have undertaken two types of tasks. One is an orderly description of all classes of variables that might be included in a complete domain of such variables. The other is to create a generalizable latent structure model of this domain, which identifies the dimensions along which the variables can be arrayed. These two tasks were attempted for both the predictor domain and the job performance domain.

The basic design of Project A is that of a multistage validation study. Its scope is large and the data collection effort is massive and included:

- *Routinely collected personnel data*: ASVAB scores, training school grades, and scores on the Skills Qualification Test (a paper-and-pencil test of current job knowledge), available from the Army's personnel records.

- *Data for construction of predictor variables*: At three points in time, versions of the predictor instruments were administered to samples representing the "preliminary," "trial," and the final "experimental" predictor batteries.
- *Criterion measurements*: Criterion job performance measurements were made (or were planned to be made) at two points in time for each of two samples.

THE ARMY'S DATA COLLECTION

Figure 1 is a summary of the Army data collections. There are essentially three samples or designs under way. The first sample is the preliminary longitudinal sample. It includes soldiers from four military occupational specialties who enlisted and were initially tested in 1983-1984. The second sample is the concurrent validation sample and includes soldiers from 19 military occupational specialties (including the initial 4) who also enlisted in 1983-1984 and who were tested in 1985. There is some overlap between these first two samples: the second includes those from the first who could be found and tested at the later date. The third sample, the longitudinal validation sample, includes soldiers from 21 military occupational specialties (including most of the 19) who enlisted in 1986-1987. These soldiers were tested with predictor variables as they entered their tour of duty and will be tested with criterion measures as they continue their military service. Each sample includes soldiers tested with both predictor and criterion measures during their first tour of duty; the second and third samples include soldiers tested also with criterion measures during their second tour of duty.

For the first sample, a preliminary or test version of the predictor instruments was administered during the project's development period (1983-1984). This was an off-the-shelf set of predictor instruments, including cognitive and noncognitive measures (interests, biographical data, and personality) selected for later comparison with project-developed tests. It was designed not to overlap with the ASVAB. Training grades and indices of early attrition were the criteria used to evaluate the variables of the preliminary predictor battery for inclusion in the next phase of the research (Personnel Decisions Research Institute, 1985; Wing, Hough, and Peterson, 1987). The overlap with the ASVAB was evaluated and judged to be not large. The sample is referred to as the preliminary longitudinal sample; it

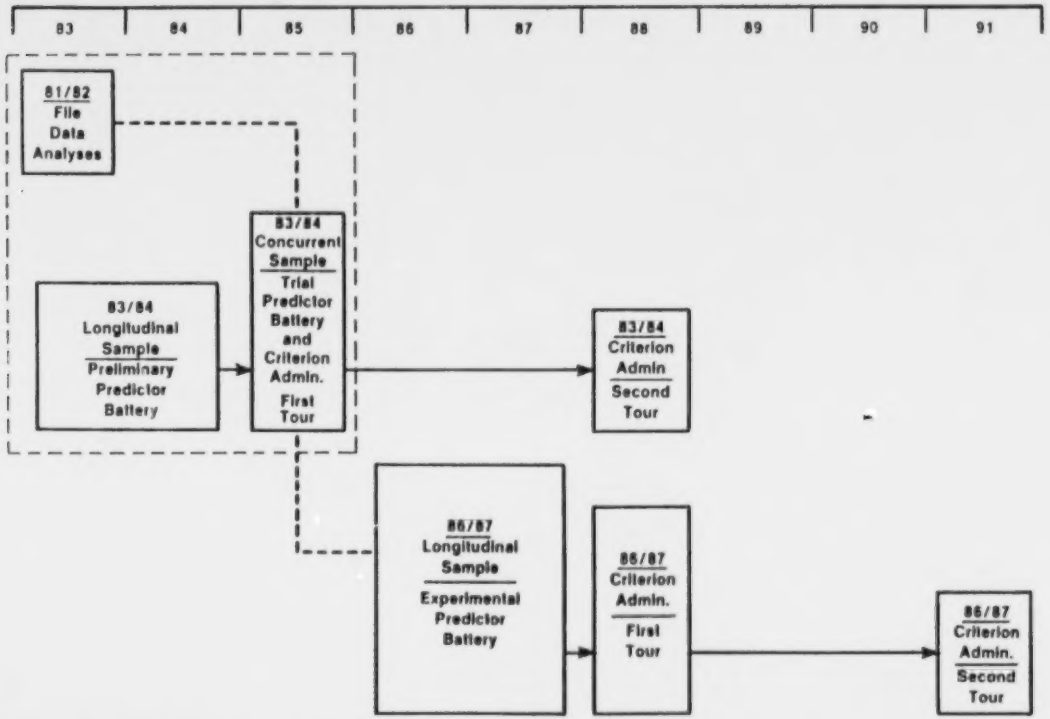

FIGURE 1 The Army's data collections
Source: Campbell (1986).

was to include about 2,000 recruits in each of 4 selected military occupational specialties.

The second, larger sample, the concurrent validation sample, was composed of 400-600 incumbents in each of 19 military occupational specialties. Each of the 4 military occupational specialties of the preliminary longitudinal sample (discussed above) provided approximately 100 of the 400-600 required incumbents for that specialty. During 1985, all incumbents were simultaneously given the trial predictor battery and criterion instruments. The trial predictor battery consisted of predictor measures derived from the preliminary battery but developed specifically for the project. It included computerized assessments as well as the cognitive and noncognitive measures represented in the preliminary battery. The criterion measures, discussed in more detail below, were also developed specifically for this project. For a subset of nine military occupational specialties, criteria included job sample tests, job knowledge tests and rating scales linked to the job sample tests, and occupation-specific behaviorally anchored rating scales. For all 19 military occupational specialties tested, criterion measures included job knowledge tests as well as ratings and self-reports of nonoccupation-specific, Army-wide behavior.

The third sample, a still larger longitudinal sample composed of over 40,000 recruits from 21 military occupational specialties, was part of the data collection scheduled for 1986-1987. A refinement of the trial battery administered to the second, concurrent validation sample but known now as the experimental battery was used, as were refinements of the criterion measures. Soldiers were tested with the experimental battery shortly after they began their Army careers. The criterion measures were administered approximately 18 months later.

It should be noted from Figure 1 that the design includes the retesting of some soldiers who reenlist for a second tour of duty. Some members of the 1983-1984 concurrent sample will be readministered a subset of the same criterion measures as well as new criterion measures during their second tour of duty, in 1988, as will some members of the 1986-1987 longitudinal sample scheduled to be retested in 1991. The testing will provide longitudinal data on the ability of the predictor variables to account for job performance during the soldiers' second tour of duty. The researchers anticipate that performance during the second tour will depend more heavily on supervisory, leadership, and higher-level technical skills.

There are scattered references in Campbell (1986) to the researchers' desire to develop models using samples of soldiers from a sample of military jobs. The sampling of soldiers and of jobs appears purposive rather than probabilistic. The selection of military occupational specialties dominated the selection of soldiers. The selection of military occupational specialties went through several iterations, to balance various characteristics such as size or fill rate, minority/gender representation, and aptitude area requirements. The final sample of military occupational specialties was then reviewed by experts and the Army's Governance Advisory Group, which was composed of general officers. Following the recommendations of this group, a final list of military occupational specialties was assembled.[1]

Soldiers in 19 military occupations were studied; for each military occupational specialty, a series of rating scales of general performance factors (e.g., giving peer leadership and support, maintaining equipment, self-discipline) and an occupation-specific written test of training achievement were administered. In addition, five performance indices were derived from administrative records.

Budget constraints, however, restricted the development of other, more comprehensive occupation-specific performance measures (hands-on job samples, written tests, and occupation-specific rating scales) to only nine military occupational specialties: infantryman, cannon crewman, armor crewman, single-channel radio operator, light wheel vehicle mechanic, motor transport operator, administrative specialist, medical specialist, and military police. These nine occupations were designated "Batch A." For the 10 remaining occupations, in place of the hands-on and other occupation-specific measures, performance ratings were obtained for 13 characteristic tasks that could (potentially) be performed by incumbents in all military occupations (e.g., first aid). These occupations, the members of "Batch Z," were: combat engineer, MANPADS crewman, TOW/Dragon repairman, carpentry/masonry specialist, chemical

[1] Staff of the contractor and officers assigned to the Army Research Institute sorted a sample of 111 job descriptions into groups on the basis of their judgments of similarities and differences in job activities. Scaling procedures were used to cluster the jobs into major groups. The intent was to ensure that the selected sample of occupations was representative of these major groups as derived from the scaling. The results of this exercise were provided to the Governance Advisory Group, which replaced two occupations in the sample (Campbell, 1986).

operations specialist, ammunition specialist, utility helicopter repairman, petroleum supply specialist, supply specialist, and food service specialist.

The selection of soldiers for the concurrent validation effort began with the selection of a set of large Army bases across the continental United States and in Europe in which one or more of the selected military occupational specialties were well represented. Dates were negotiated for the data collection efforts, and requests were made of each base for certain numbers of soldiers in certain military occupational specialties who had entered the Army for the first time during 1983 or 1984. Not all military occupational specialties were tested at each base. Race/ethnicity and gender distributions were specified; the base was given the responsibility for selecting soldiers who met the above requirements to be available for a day or two of testing. Soldiers in Batch A occupations required two days of testing; those in Batch Z occupations required only one day. Encouragement was given to make these samples "representative." The only exception to the above procedure involved soldiers in the 4 military occupational specialties (all in Batch A) who had been administered the preliminary predictor battery some 18 months before. Those individuals were located and, if resident at one of the bases where testing for that military occupational specialty was taking place, were requested by name to be part of the concurrent validation sample.

Predictor Development

The range of tests and samples used in the early stages of this research reflects the amount of effort put into instrument development. The procedure used to develop the predictors of job performance is described as "somewhat unique" (Campbell, 1986). It is said that the traditional procedure uses job analyses that permit experts to identify the types of knowledge, skills, and abilities that would be required for successful performance;[2] instruments are then developed to measure those factors.

The project researchers used a different strategy. They began by conducting a broad search for predictor variables that would be appropriate for the variety of military occupations. Subsequently

[2]The researchers note that an additional step is usually required, which is identifying the factors that are trainable versus those that must be selected for.

they planned to select the "best bet" set of variables, develop adequate measures of them, and then submit the measures to progressive refinement via a succession of operational tests. This strategy was intended to develop a single predictor battery that could be used across the entire range of military occupations—rather than custom-made batteries for specific occupations (or groups of occupations).

The researchers began with an extensive search of the personnel literature for possible predictor constructs. Several hundred constructs were identified during this first stage. Successive rounds of internal reviews reduced the list to 53 constructs that were thought to be most potentially useful. The aim was to select variables that would best supplement the ASVAB in predicting job performance across the range of military occupations. Table 1 presents an overview of the predictor data. Project scientists followed an analogous route to identify a set of 72 performance factors (see Wing, Peterson, and Hoffman, 1984). Expert judges provided estimates of the validity of each predictor construct for each performance factor. These estimates were used in determining the set of predictor variables to be used in the successive versions of the predictor battery.[3]

Job Performance Criteria

Contrary to the Joint-Service Project focus on job proficiency, the goal of performance measurement in the Army's Project A was to build a model of the total performance domain, identify the major factors or dimensions of this domain, and then create reliable, valid measures for each factor (Campbell, 1986). Three types of measurements were used: (1) hands-on tests of job performance, (2) multiple-choice tests of job knowledge, and (3) ratings of job performance.

For practical and theoretical reasons, the researchers restricted their focus to individual performance rather than group performance.

[3]A total of 35 experts in personnel selection were asked to estimate the correlations between each of the 53 predictors and each of the 72 criterion measures as corrected for range restriction and criterion attenuation. The resultant 53 by 72 matrix of estimates was first evaluated for interjudge agreement, or reliability, which was high. The estimates were then factor analyzed, by rows (predictors) and by columns (criteria). Clusters and factors of predictor measures were found as well as clusters and factors of criterion measures (Wing, Peterson, and Hoffman, 1984). The contents of the preliminary battery (and the subsequent trial and experimental batteries) were based on the clusters of predictor variables determined in this analysis.

Soldiers primarily are recruited into and are moved about in the Army as individuals. Project A was designed to focus on individual differences, in predictors and on performance measures, and to evaluate the relationships between predictors and criteria for a large variety of very different individuals. An attempt was made to obtain available group performance data, but doing so in any systematic way proved to be impossible. Finally, the high likelihood of different kinds of interaction among and between individuals in any given group makes prediction of unit performance a particularly intractable problem at the present time. Research on this topic is comparatively scarce.

The researchers conceived of job performance as multifaceted. They assumed that each job was composed of both elements unique to that job and elements shared by all jobs in the Army. For job skill proficiency—"can do" aspects of performance in the Batch A occupations—the researchers posited two classes of factors for entry-level military occupations. The first is occupation-specific (e.g., performance of tasks that are unique to a given occupation), and the second is common across occupations. The common or Army-wide tasks would not differentiate among occupations.

Going beyond the scope of the Joint-Service Project, the Army research also looked at the "does do" aspects of performance (e.g., discipline, teamwork). These were presumed to be common to all Army occupations. Given this perspective, the researchers attempted to model the latent structure of job performance measurements, although they noted the difficulties involved.

The primary Project A approach to job analysis is a task-based approach. For each Batch A military occupational specialty, the researchers used the *Soldier's Manual* and the results of Army task description surveys to identify the tasks to be used in constructing their performance measures. The researchers noted the differing levels of descriptive specificity employed in these two sources, and the artificial way in which behaviors were sometimes divided into tasks. To reduce the resultant ambiguity, the researchers selected statements that "described a complete operation, which had a recognizable beginning and end, and which were relatively independent of other tasks" (Campbell, 1986:13). After several iterations and reviews by experts in the occupations being described, the researchers obtained lists of 130-180 tasks for each Batch A military occupational specialty.

TABLE 1 Hierarchical Map of Predictor Space

Constructs	Clusters	Factors
1. Verbal comprehension 5. Reading comprehension 16. Ideational fluency 18. Analogical reasoning 21. Omnibus intelligence/aptitude 22. Word fluency	A. Verbal ability/ general intelligence	
4. Word problems 8. Inductive reasoning concept formation 10. Deductive logic	B. Reasoning	
2. Numerical computation 3. Use of formula/number problems	C. Number ability	COGNITIVE ABILITIES
12. Perceptual speed and accuracy	N. Perceptual speed and accuracy	
49. Investigative interests	U. Investigative interests	
14. Rote memory 17. Follow directions	J. Memory	
19. Figural reasoning 23. Verbal and figural closure	F. Closure	
6. Two-dimensional mental rotation 7. Three-dimensional mental rotation 9. Spatial visualization 11. Field dependence (negative) 15. Place memory (visual memory) 20. Spatial scanning	E. Visualization/spatial	VISUALIZATION/ SPATIAL
24. Processing efficiency 25. Selective attention 26. Time sharing	G. Mental information processing	INFORMATION PROCESSING
13. Mechanical comprehension	L. Mechanical comprehension	
48. Realistic interests 51. Artistic interests (negative)	M. Realistic vs. artistic interests	MECHANICAL
28. Control precision 29. Rate control 32. Arm-hand steadiness 34. Aiming	I. Steadiness/precision	
27. Multilimb coordination 35. Speed of arm movement	D. Coordination	PSYCHOMOTOR
30. Manual dexterity 31. Finger dexterity 33. Wrist-finger speed	K. Dexterity	

BEST COPY AVAILABLE

TABLE 1 (continued)

Constructs	Clusters	Factors
39. Sociability 52. Social interests	Q. Sociability	SOCIAL SKILLS
50. Enterprising interests	R. Enterprising interests	
36. Involvement in athletics and physical conditioning 37. Energy level	T. Athletic abilities/energy	VIGOR
41. Dominance 42. Self-esteem	S. Dominance/self-esteem	
40. Traditional values 43. Conscientiousness 46. Nondelinquency 53. Conventional interests	N. Traditional values/ conventionality/ nondelinquency	
44. Locus of control 47. Work orientation	O. Work orientation/locus of control	MOTIVATION/ STABILITY
38. Cooperativeness 45. Emotional stability	P. Cooperation/emotional stability	

Source: Campbell (1986).

There were two steps in the selection of tasks to be tested. First, additional expert judgments were used to determine task difficulty and importance and to group tasks based on their similarity of content. The sampling of tasks from these groups, to be measured in the job performance tests, was done with a modified Delphi approach (Campbell, 1986:13).

> Each member of a team of task selectors was asked to select 30 tasks from the population of tasks such that the selected tasks were representative of . . . [job] . . . content, were important, and represented a range of difficulty. The individual judge's choices were then regressed on the task characteristics and both the choices and the captured "policy" of each person were fed back to the group members, who each revised their choices as they saw fit.

It is stated that this procedure quickly achieved consensus on the tasks to be selected and that the resultant selections were thoroughly reviewed by the Army command responsible for those jobs.

Second, for each Batch A occupation, 15 of the 30 tasks were selected for inclusion in the hands-on job tests. The researchers then developed hands-on job samples and numerical ratings scales for these 15 tasks, paper-and-pencil job knowledge tests for all 30

tasks, and a job history questionnaire. The hands-on job performance measures broke down the 15 tasks into a number of steps that were to be scored on a pass-fail basis. Each job knowledge test, for each Batch A occupation, was to have one or more items per task, a requirement met for nearly all of the 270 tasks (9 jobs times 30 tasks) (Rumsey, Osborn, and Ford, 1985).

Project A also experimented with an approach to describing the important elements of occupations, called the "critical incident method." It leads to a different conception of job performance. The important characteristic of this method is that it begins with incumbents or subject matter experts being asked to write down short but compiete descriptions of individual performance on their job, which was either extraordinarily good or extraordinarily poor. These descriptions are reviewed by measurement experts and sorted into categories or dimensions on the basis of the similarity of the demonstrated behavior. The structure and content of these categories is typically confirmed by incumbents. There are many uses for critical incident methodology besides performance assessment, such as equipment design (or misdesign). In performance evaluation, some of the incidents can be used as anchor points for scales to evaluate individuals on the dimension. Typically, a job might be covered by 8 to 12 critical incident scales; this coverage includes those areas in which people perform differently but not necessarily those areas in which people spend the most time. The critical incident method is particularly useful in validation research as it identifies the extremes of behavior that distinguish the superior from the inferior performers. It does not, however, necessarily identify representative samples of job behaviors.

Using the critical incident methodology, two groups of rating scales were developed. The first group consisted of a set of rating scales for each of the Batch A occupations. The second group consisted of one set of rating scales, for elements common to all Batch A and Batch Z occupations. There were two parallel tracks followed in this phase of the research, to obtain the two groups of rating scales. For both tracks, panels of noncommissioned and commissioned officers generated thousands of descriptions of extremely good or extremely poor soldier performance. Panels in the first track, from the Batch A occupations, were asked to focus on critical incidents that were specific to their own particular military occupational specialty. Each panel consisted of members of the same military occupation. Panels in the second track, from both Batch A and Batch

Z occupations, were asked to focus on incidents that could potentially occur in any military occupation. Behavioral summary rating scales were then developed for each of the Batch A occupations and for Army occupations in general. There were several scales specific to each of the 9 Batch A specialties[4] as well as 10 Army-wide scales. There were three additional Army-wide rating scales: overall job performance, noncommissioned officer potential, and expected combat performance. The last consisted of 40 separate rating scales and was behavioral expectation rather than behavioral summary.

The ratings scales for the 13 tasks representative of all Army occupations, to be used only for Batch Z occupations, were similarly developed using incumbents from the Batch Z occupations. These ratings were not used in any of the analysis discussed below.

In addition to the measures described above, a training knowledge test was also developed for each of the 19 Batch A and Batch Z occupations by writing items to describe those tasks that are both taught in training and performed on the job. These training achievement tests were not directly related to the subset of 15 or 30 tasks tested for the Batch A military occupational specialties.

Finally, an attempt was made to use archival records to provide performance data. The researchers thought that the 201 file (Military Personnel Records Jacket) would be the most promising source of information; however, it was the most expensive to search since it must be done by hand. The soldier's 201 file tends to be out-of-date because it takes time for records to be entered into the file and because the file lags in following a soldier from assignment to assignment. Because the 201 file proved to be an awkward source of data, a self-report form was developed for soldiers to list their recall of disciplinary actions, letters of commendation, etc. A verification study of the self-reports of 500 soldiers against their 201 files indicated that their self-reports contained a greater number of both positive and negative incidents than did the archival records. The verification was generally successful, and four performance indicators were developed for the self-report form. A fifth performance indicator, promotion rate, was developed from computerized records.

[4]The number of scales varied from specialty to specialty; there could be as few as 6 or as many as 12.

Missing Data

As with all data gathering programs, complete data sometimes cannot be collected. Hands-on test administrations pose particular difficulties. The Project A researchers report, for example, that on one occasion they were caught by a surprise rainstorm that caused a nearby creek to overflow (Wise, McHenry, and Young, 1986). The flood carried off some of the test data and interfered with the completion of testing. In addition to acts of God, incomplete data gathering also occurred because of inadequacies in the training of personnel assigned to administer or score the tests. The researchers reported occasions on which a scorer apparently was unable to follow the directions printed on the score sheet, so one or more steps of the tested task were not scored. For the majority of soldiers tested (3,370 of 5,268), data were missing for at least one hands-on task. In 612 of 5,268 cases, more than 10 percent of the data from the hands-on tests was missing. It is said that these relatively large numbers of missing data occurred primarily because the equipment varied from site to site or broke down. Some tasks required specific items of equipment that were not available at all the bases where a given military specialty was tested.

To deal with missing data, the Project A researchers:

- Treated an instrument as missing entirely if 10 percent of constituent items or scales were missing. (In the case of the hands-on tests, this cutoff was set at 15 percent.)
- Computed correlations between a subject's score on particular items and the "easiness" of the item (defined as the proportion of subjects in the sample who answered the question correctly). When there was no relationship (i.e., a zero correlation) between scores and easiness, the researchers assumed that the subject was responding randomly, and the subject's score on the instrument was treated as missing.
- Imputed scores for missing items for instruments that were not defined as missing entirely,[5] and calculated total scores on the instrument. This imputation was done for the not-reached items of the written tests.
- Computed scores on performance constructs (described below) from the constituent component scores. If one or more

[5]Imputations were done using a procedure developed for the Center for Educational Statistics; see Wise and McLaughlin (1980).

component scores were missing (or treated as missing by the above rules), imputed performance "construct" scores as long as no more than half the component scores (included in that construct) were missing.

Performance Construct Development

The performance data collected from each soldier are summarized in Table 2. The researchers noted that the rating scales, major predictor subscales, and aggregate task-level measurement provide approximately 270 criterion scores for use in the analysis (Campbell, 1986; McHenry et al., 1986; Wise et al., 1986). Faced with this array of variables, the researchers undertook an initial stage of data reduction--opting to "reduce collinearity as much as possible and deal at the construct level" (Campbell, 1986:16). Two reasons motivated this decision: the researchers wished to generalize their findings broadly, and they were concerned about the potential for capitalizing on chance relationships among so many variables in such a large data base. In addition, the researchers noted that Project A is intended to estimate a variety of predictor-criterion relationships, and they worried that it might be impossible to detect instances of differential prediction with sufficient statistical precision among such a large number of variables. This led the researchers to reduce the number of parameters to be estimated in order to evaluate the underlying structure of the total predictor-performance domain. They limited their analyses to the nine Batch A occupations.

Initial data reduction was achieved through a mixture of factor scalings, a priori groupings of measurements, and expert judgment. Because occupation-specific, individual task (item) results from the hands-on, job knowledge, and school knowledge tests were not as internally consistent as had been hoped, the researchers grouped the task scores for each of the Batch A military occupational specialties into functional categories based on similarity of task content. The results were 8 to 15 functional categories per job. For each of the three types of measures for each job, aggregate scores were computed for each of the functional categories.

These aggregate scores were then factor analyzed (separately within occupation and type of test). The analyses yielded two results that were common across the different jobs and the three measurements: Measures of occupation-specific task performance loaded on different factors than the measures of performance on general tasks

TABLE 2 Performance-Criterion Measures Used in Concurrent Validation Samples

---

Performance Measures Common to Batch A and Batch Z MOS (Jobs)

1. Ten behaviorally anchored rating scales designed to measure factors of non-job-specific performance (e.g., giving peer leadership and support, maintaining equipment, self-discipline).

2. Single scale rating of overall job performance.

3. Single scale rating of NCO (noncommissioned officer) potential.

4. Paper-and-pencil Test of Training Achievement developed for each of the 19 MOS (130-210 items each).

5. A 40-item summated rating scale for the assessment of expected combat performance.

6. Five performance indicators from administrative records. The first four are obtained via self-report and the last one from computerised records:

   - o Total number of awards and letters of commendation;
   - o Physical fitness qualification;
   - o Number of disciplinary infractions;
   - o Rifle markmanship qualification score; and
   - o Promotion rate (in deviation units).

Performance Measures for Batch A Only

7. Job-sample (hands-on) test of MOS-specific task proficiency.

   - o Individual is tested on each of 15 major job tasks.

8. Paper-and-pencil job knowledge tests designed to measure task-specific job knowledge.

   - o Individual is scored on 150-200 multiple-choice items representing 30 major job tasks; 15 of the tasks were also measured hands-on.

9. Rating scale measures of specific task performance on the 15 tasks also measured with the knowledge tests and the hands-on measures.

10. MOS-specific behaviorally anchored ratings scales. From 7 to 13 BARS were developed for each MOS to represent the major factors that constituted job-specific technical and task proficiency.

Performance Measures for Batch Z Only

11. Ratings of performance on 13 representative "common" tasks. The Army specifies a series of common tasks (e.g., several first aid tasks) that everyone should be able to perform.

TABLE 2 (continued)

Auxiliary Measures Included in Criterion Battery
12. Job History Questionnaire, which asks for information about frequency and recency of performance of the MOS-specific tasks.
13. Work Environment Description Questionnaire--a 141-item questionnaire assessing situational/environmental characteristics, leadership climate, and reward preferences.

Note: All rating measures were obtained from approximately 2 supervisors and 3 peers for each person rated.

Source: Campbell (1986).

common to all military occupations. In addition, a fairly similar set of factors emerged for the tasks common to all military occupational specialties. These similarities held across the nine different military occupational specialties and the three types of measures.

The factor analyses further led the researchers to define content categories of job performance. These categories were said to reflect theoretical concerns rather than a strict adherence to the empirical results. For the functional categories across all occupations, from the hands-on, job knowledge tests, and school knowledge tests, six content categories were identified:

(1) Basic soldiering skills (weapons, navigation, field techniques, customs and laws);
(2) Safety and survival (first aid, nuclear-biological-chemical safety);
(3) Communications (radio operation);
(4) Vehicle maintenance;
(5) Identification of friendly and enemy aircraft and vehicles; and
(6) Technical skills specific to a particular job.

The ratings provided by supervisors and peers for each of the 15 hands-on tasks for each military occupational specialty, were insufficiently reliable for subsequent analysis. This was primarily due to missing data: the raters had "insufficient opportunity to observe" the soldier performing the task.

In a subsequent analysis, the 10 individual performance, Army-wide, behaviorally anchored rating scales used for both Batch A and

Batch Z military occupational specialties, combined with the noncommissioned officer potential rating scale, were then factor analyzed to produce another set of performance scores. These analyses yielded three additional factors:

(7) Effort/leadership (competence and effort in job performance, leadership, self-development);
(8) Self-discipline (integrity, self-control, ability to follow regulations); and
(9) Fitness and appearance (physical fitness, proper military bearing).

Two additional performance scores were then derived from the occupation-specific behaviorally anchored rating scales, across the Batch A military occupational specialties, based on the results of factor analyses and expert judgment:

(10) Technical content specific and central to performance in the given occupation and
(11) Remaining, less central aspects of performance in the given occupation.

Next, the 40 items in the rating scale of expected combat performance were factor analyzed, yielding two more factors:

(12) Exemplary effort, skill, or courage under stressful conditions and
(13) Failure to follow instructions, lack of discipline under stressful conditions.

Finally, six additional measures were also used, five from the administrative records and the sixth being the overall performance rating:

(14) Letters and certificates;
(15) Physical readiness test score;
(16) M16 qualification score;
(17) Articles 15/flag actions;
(18) Promotion rate deviation score; and
(19) Overall performance rating.

The final array of summary performance variables resulting from this data reduction is as follows: for the first six factors there were 2-5 content category scores from hands-on job samples, 2-6 content category scores from job knowledge tests, and 2-6 content category scores from school knowledge tests; this variation in number of scores occurred across the Batch A occupations (Factors 1-6). For attributes

required by all military occupations, there were the 3 rating factors derived from the 10 Army-wide behaviorally anchored rating scales (Factors 7-9; the leadership potential rating was dropped from these and subsequent analyses). From the Batch A occupation-specific behaviorally anchored rating scales, there were 2 rating factors for tasks specific to a particular military occupation (Factors 10-11). Then there were the 2 expected combat performance factors (Factors 12-13). Finally, there was an overall effectiveness rating and the 5 archival measures of performance derived from administrative records or self-reports (Factors 14-19), for a minimum of 19 and a maximum of 30 variables for each soldier in each military occupation (McHenry et al., 1986; Wise et al., 1986).

THE ARMY'S ANALYSIS

Latent Structure of Performance

The researchers began the process of fitting the data to and modeling the structure of the complete predictor/performance domain (Campbell, 1986; McHenry et al., 1986; Wise et al., 1986). Their analytic goal was to translate the specifications of their initial model into matrices of parameter estimates that would be input to the computational/theoretical procedures of LISREL. LISREL evaluates the fit of data to models positing causal or asymmetric relationships among variables, such as performance variables being predicted by variables in a predictor battery, but not vice versa. The initial target model they employed is shown in Figure 2.

The initial step was to refactor the 19-30 summary job performance criterion measures for each soldier for each of the Batch A occupations. There were three important findings: first and of primary importance was the emergence of two methods factors, one for ratings and one for written tests. There were insufficient measures for methods factors to emerge for hands-on tests and administrative measures. Second, the Army-wide ratings lined up with the administrative measures but not with the job skills measures (hands-on and written tests). The job-specific ratings appeared to be measuring both technical knowledge and skill as well as effort and leadership, although emphasizing effort and leadership. Third, and most important, the job performance content category measures did not "cross" methods: the job knowledge and school knowledge test scores were mostly unrelated to the hands-on test scores. From these findings the

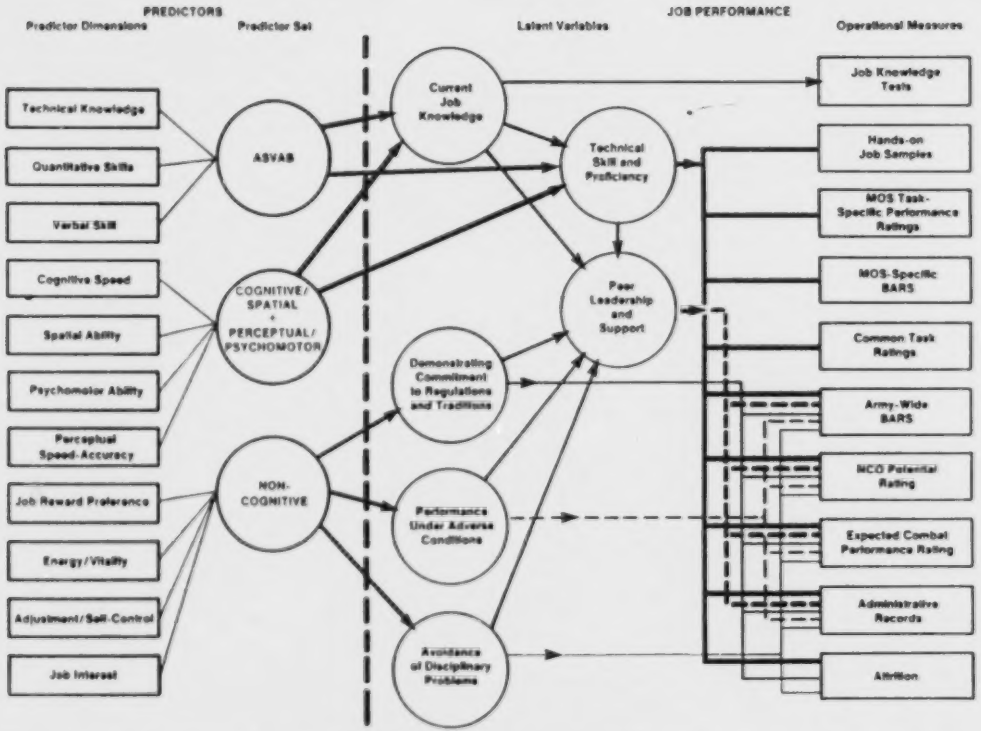

FIGURE 2 A proposed structural model of job performance
Source: Campbell (1986)

researchers constructed a revised model of performance that included five performance factors and two methods factors. (A sixth performance factor included only one administrative records measure, the M16 qualifications score, and consequently was dropped from further analyses.) The five substantive performance constructs are described in Table 3.

The next step was to refactor the variables within each of the Batch A military specialties using LISREL to confirm the performance model. The model was constrained such that correlations between each methods factor and each performance construct were zero. It is important to note that this use of LISREL analytic procedures to confirm a model, using the same data that were used to construct the model, could limit the resulting findings by overemphasizing chance effects. The initial pass indicated problems with parameter estimates for several jobs. One set of these was reestimated, using the same data, and resubmitted to LISREL. Despite the violation of assumptions, the data appeared to fit the model quite well: only two of the nine chi-squares were statistically significant ($p < .05$) misfits.

The final step was to use the LISREL multigroups option to confirm the model across occupations, using jobs as replications. There were again potential problems. One was the differential selectivity of the jobs. The jobs had different entrance requirements for various types of ability, which could lead to different values for reliabilities and for validities of predictor-criterion relationships. The other was that this model required the same number of observed variables for each job. As noted above, there were missing data across the 9 military specialties among the 18 performance skill content categories (3 types of measures times 6 content categories). Not all occupations had each class of content for each type of measures. Some of the estimated intercorrelations had to be set to constants to eliminate subsequent analysis problems; that is, the input parameters as well as the degrees of freedom were adjusted to deal with the missing data problem; no attempt was made to correct for any occupation-specific differential selectivity. The LISREL analysis showed the model to provide an adequate fit ($p > .05$) to the data.

A brief recapitulation of the analyses is as follows: First, most of the performance scores (with missing data imputed) were factor-analyzed within one of two categories, job skills or ratings, to develop new variables. These new variables were then created for each soldier and refactored within occupation by LISREL, adjusted for parameter

TABLE 3 Performance Factors Representing the Common Latent Structure Across All Jobs in the Project A Sample

1. Task proficiency: specific core technical skills: The proficiency with which the individual performs the tasks that are "central" to his or her job (MOS). The tasks represent the core of the job; they are the primary definers from job to job.

- The subscales representing core content in both the knowledge tests and the job sample tests that loaded on this factor were summed, standardized, and then added together for a total factor score. The factor score does not include any rating measures.

2. Task proficiency: general or common skills: In addition to the core technical content specific to an MOS, individuals in every MOS are responsible for being able to perform a variety of general or common tasks--e.g., use of basic weapons, first aid, etc. This factor represents proficiency on these general tasks.

- The same procedure (as for factor one) was used to compute the knowledge and hands-on general task scores, standardized within methods, and with the two standardized scores added together.

3. Peer leadership, effort, and self-development: Reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient?

- Five scales from the Army-wide Behaviorally Anchored Rating Scale (BARS) rating form (Technical Knowledge/Skill, Leadership, Effort, Self-Development, and Maintaining Assigned Equipment), the expected combat performance scales, the job-specific BARS scales, the general performance rating, and the total number of commendations and awards received by the individual were summed for this factor.

4. Maintaining personal discipline: Reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates responsibility in day-to-day behavior, and does not create disciplinary problems.

- Scores on this factor are composed of three Army-wide BARS scales (Following Regulations, Self-Control, and Integrity) and two indices from the administrative records (number of disciplinary actions and promotion rate).

5. Physical fitness and military bearing: Represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

- Factor scores are the sum of the physical fitness qualification score from the individual's personnel record and two rating scales from the Army-wide BARS (Military Appearance and Physical Fitness).

Note: The criterion measures that comprise each factor are as indicated.

Source: Campbell (1986).

estimation problems by reanalysis of the data, and factored again by LISREL. Finally, these adjusted variables, along with constants for missing values, were analyzed again by LISREL to confirm the overall model produced during the second phase. The investigators did not rely solely on results of data analyses to build and verify their model. However, repeated similar analyses on the same data set may lead to problems in the interpretation and generalization of results.

Differential Validity

The investigators subsequently reported a series of analyses of the differential validity of various combinations of predictors for different military occupations. There is, of course, substantial variation in the associations observed between predictors (e.g., the ASVAB) and different performance criteria. For example, the various ASVAB composites correlate .49 with hands-on job test results in the nine jobs studied, but (in accord with reasonable expectations) its correlation with the personal discipline performance factor is only .20.

Averaging across military occupational specialties, the investigators found that the predictors they tested added relatively little to the ability of the ASVAB to predict (a) total score on the hands-on tests, (b) scores on the "core technical" tasks, (c) general soldiering performance, or (d) scores on the effort/leadership factor.[6] Exceptions occurred for personal discipline and fitness/military bearing. For these performance criteria, inclusion of the temperament and biographical factors increased the multiple correlation from .21 (ASVAB alone) to .35 (discipline) and .40 (fitness/military bearing).[7]

The investigators reported stepwise regressions for subscales of these predictors. It appears from these regressions that the physical fitness predictor subscale plays the major role in increasing the predictability (over predictions using only ASVAB) of the physical

[6]For hands-on tests, the multiple correlation with ASVAB was .49 compared with .53 when all trial battery predictors were used. For core technical proficiency tasks, the multiple correlation with ASVAB was .48 compared with .52 when all trial battery predictors were used. For general soldiering performance, the multiple correlation with ASVAB was .51 compared with .55 when all trial battery predictors were used. For effort/leadership, the multiple correlation with ASVAB was .46 compared with .53 when all trial battery predictors were used. (In each case, a measurement methods factor was partialled out of the correlations.)

[7]Here again the criterion scores were adjusted by removing the linear effect of the measurement methods factor.

fitness/military bearing criterion. For personal discipline, the major contribution appears to result from the dependability subscale of the temperament predictor.

The investigators also reported the results of stepwise regressions using the predictor subscales *within* each of the military occupations. As Table 4 shows, the occupation-specific core technical proficiency criteria show noteworthy variations in association with the predictors. For example, the occupations of cannon crewman (13B) and military police (95B) show relatively low levels of association with the ASVAB predictors, while infantryman (11B) and light wheel vehicle mechanic (63B) show considerably higher levels of association with the ASVAB predictors.[8]

These coefficients do not appear to have been corrected for range restriction or criterion unreliability. The true differences might be smaller or larger. It will be more informative to compare the coefficients after the appropriate corrections have been made

Weighting Constructs

Empirical studies (Sadacca, de Vera, and DiFazio, 1986) were also made of alternative procedures for assigning weights to performance dimensions so that these weights reflect expert opinion on how important each performance dimension is to the evaluation of overall performance. All of the procedures elicit judgments from experts about the relative importance of various dimensions. Three initial experiments were conducted on weighting job performance dimensions to select one or more of these alternative procedures that would both be acceptable to the Army and would produce a reliable, valid set of weights for each of the 19 military occupations.

On the basis of these initial experiments, two methods of weighting were selected for further research: direct estimation and conjoint paired comparison ranking. In each, the five performance constructs evaluated were those determined in the performance construct development phase, described above:

(1) Maintaining personal discipline;
(2) Military bearing/appearance and physical fitness;
(3) Exercise of leadership, effort, and self-development;

[8] For the vehicle mechanic occupation (63B), this prediction results entirely from the technical subscale of the ASVAB.

TABLE 4 Results of Stepwise Regressions for MOS-Specific Core Technical Proficiency for Each of the 9 Batch A Occupations

Predictor Constructs	MOS 11B	13B	19E	31C	63B	64C	71L	91A	95B
ASVAB FACTORS									
Verbal	0.17	--	0.10	0.21	--	--	0.08	0.26	0.13
Quantitative	0.09	--	--	0.30	--	--	0.27	--	--
Technical	0.10	--	0.16	--	0.35	0.30	-0.13	0.12	--
Speed	--	--	--	--	--	-0.07	--	0.13	--
SPATIAL									
Overall Spatial	0.20	0.25	0.19	--	0.14	0.16	0.25	0.23	0.22
COMPUTER									
Complex Perc Speed	-0.18	--	--	--	--	-0.12	--	--	--
Complex Perc Accy	0.13	--	0.09	0.10	--	0.14	0.15	--	0.09
Number Speed/Accy	--	--	-0.09	--	--	--	--	--	-0.11
Psychomotor	--	--	--	--	--	--	--	--	--
Simp Reaction Accy	--	--	0.07	--	--	--	--	--	--
Simp Reaction Speed	--	-0.10	--	--	-0.11	--	--	--	--
TEMPERAMENT									
Adjustment	-0.08	--	--	-0.09	--	--	--	--	--
Dependability	0.12	--	0.10	0.15	0.13	0.07	0.11	0.22	0.12
Surgency	--	--	--	--	--	--	--	--	--
Phys Condition	--	--	-0.09	--	-0.06	--	--	-0.13	--
INTERESTS									
Combat	0.15	0.21	0.17	--	--	--	--	0.16	--
Machines	--	--	--	0.21	0.32	--	-0.16	--	--
Audiovisual	--	--	--	--	-0.14	--	--	-0.09	-0.13
Technical	--	--	--	--	--	--	0.12	--	--
Food Service	-0.07	--	--	--	--	--	--	--	--
Protective Svc	--	-0.08	--	--	-0.08	--	--	--	--
JOB PREFERENCES									
Support	--	--	--	--	--	0.09	--	0.12	0.09
Autonomy	--	0.09	--	-0.11	--	--	--	--	--
Routine	-0.06	-0.11	--	--	--	--	--	0.07	--
ADJUSTED, UNCORRECTED R	0.560	0.305	0.464	0.352	0.591	0.401	0.481	0.507	0.294

Note: -- indicates no data.

Source: Campbell (1986).

(4) Task proficiency: MOS-specific technical skills;
(5) Task proficiency: general soldiering skills.

In the direct estimation method, the judges were asked to rank order the five performance constructs or factors and then to assign points, assuming that the top-ranked factor was to be given a weight of 100 points. In the conjoint paired comparison method, judges were given scores on two of the five performance factors at a time, for a set

of 15 soldiers. They were required to rank order the soldiers' overall performance on the basis of their scores on these two factors. This procedure was followed for the 10 different pairs selected from the five performance factors.

Judgments were to be elicited from members of all 19 Batch A and Batch Z military occupations selected for detailed study in Project A. Judgments were sought for one scenario: a "worldwide heightened tensions" scenario.[9] Preliminary results of this study were available from rating workshops conducted with subject matter experts from 10 occupations, 5 from Batch A and 5 from Batch Z. The judges or subject matter experts included 10 groups of officers and 10 groups of noncommissioned officers for a total of 164 judges. For each military occupational specialty, the overall performance ranking for each pair of performance constructs used in the conjoint paired comparison measurement procedure were regressed against the 15 hypothetical soldiers' scores, for each judge. The investigators report that for 22 of the judges, the performance factor regression weights indicated that better (i.e., lower) overall performance rankings were given those fictitious soldiers who were less favorably evaluated on one (or more) of the performance factors. The investigators report adjusting these weights to remove this anomaly.[10]

Interjudge reliabilities were lower than expected, particularly for the noncommissioned officers. The average across the 10 military occupational specialties (and 2 rating methods) of the correlations between weights given to the performance factors by noncommissioned officers and officers was less than .50. Indeed, two occupations had negative correlations between officer and noncommissioned officer weightings, under each rating method.[11]

---

[9]Judges were instructed: "As the weights you assign may be a function of the particular context in which the soldiers' performance is being evaluated, please assume the following military situation prevails: The world is in a period of heightened tensions. There is an increasing probability that hostilities will break out. . . . The U.S. Army's mission is to support U.S. treaty obligations and to help defend the borders of allied and friendly nations. . . . U.S. Army training and other preparatory activities have been substantially increased. Most combat and associated support units are participating in frequent field exercises. Most units are being actively resupplied" (Sadacca, de Vera, and DiFazio, 1986:9).

[10]The procedure had the effect of setting the anomalous weights to (almost) zero.

[11]E.g., using the conjoint paired estimation method, a correlation of −0.55 was found between mean weights assigned by officers and noncommissioned officers to the performance factors for administrative specialist.

Two other general findings are reported: (1) the weighting of factors varied across military occupational specialties and (2) occupation-specific technical skills were usually judged to be the most important factor in job performance, and military bearing/physical fitness the least important.

THE ARMY'S CONCLUSIONS

The Army researchers drew six general conclusions from their analyses. First, job performance is multidimensional. It includes skills specific to the job and skills general to the organization, as well as personality traits and attitudes toward the job and the organization that energize the skills into actual performance. Second, the method of rating scales produces measures that reflect both proficiency and reliability. A soldier's skill, and the consistency with which he or she performs the job, will be reflected in ratings.

Third, in addition to predicting training performance, the ASVAB predicts job performance. It is a valid predictor of the "can do" or skill components of these two major aspects of performance, which are required during a soldier's first tour of duty. Fourth, all predictors are not equally valid for the different aspects of job performance. The "does do" or motivational components of performance are not well predicted by the ASVAB. Fifth, such components are validly predicted by temperament measures and biographical data. Such measures need to be included in a predictor battery if correlates of the motivational components of job performance are to be found. Finally, the primary source for differential prediction across jobs arises for those parts of performance that are specific to particular jobs. The technical skills unique to one job will be predicted by a different set of measures than those skills unique to another job.

In the future, the researchers intend to obtain estimates of validity generalization across jobs as a function of the content of the variables included in the predictor and criterion batteries. They also intend to study differences in prediction across race and gender, to estimate classification efficiency, and to estimate selection validity as a function of the weights used in criterion components and the composition of the predictor batteries used.

COMMITTEE DISCUSSION

The committee was impressed by the Army's extensive data base as well as with the uses to which the Project A scientists put it—and

this, of course, is just a beginning. There are many other useful analyses permitted by this data base. A large number of additional analyses will be undertaken during the course of Project A.

The Army scientists demonstrated that measures can be developed that are psychometrically sound and assess important aspects of job performance. They have also shown that the ASVAB and a number of other predictor measures are related to these criterion measures. The next goal of the Joint-Service Project is to move beyond the traditional validation study to produce performance information useful for setting enlistment standards. This means investing performance scores with some absolute meaning in terms of job mastery so that predictor scores not only rank order subjects by ability but tell approximately how well people at each level of the ability are likely to perform the job. Plans call for subsequent analyses of the Joint-Service Project results in modeling force allocation and force quality distributions.

Although the workshop concentrated on the Army's Project A and its preliminary results, the committee's focus and the following comments are limited to those issues, highlighted by the Army's early work, that pertain to the Joint-Service Job Performance Measurement/Enlistment Standards Project. They are to be taken as suggestions for all the Services to consider as they proceed with data collection and analyses.

Confounding Variables

- It is always better to collect data properly in the first place, although this can sometimes be difficult. Techniques for imputing missing data must be used with great care. While there will always be acts of God and equipment breakdown, attention must be paid to adequate training and supervision of test administrators and raters/scorers. Inspection of data immediately upon collection, as done in Project A, is an additional step contributing to minimizing missing data.
- The job experience of the incumbents should be taken into account in a clearly specified and objective manner. This could be time in the Service, months on the job, or some other measure. The soldiers tested entered the Army at different points in time and have different amounts of time on the job. Analyses of criterion data need to isolate the effect of experience so as to evaluate properly the role of ability in job performance.

- Any method variance shared by predictors and criteria can lead to spuriously high correlations. Correlations among instruments of the same type should be cautiously interpreted. Project A uses self-report scales, ratings, and paper-and-pencil tests, both to predict and to measure job performance.
- In many of the Project A criterion development and data analysis procedures, judgments were obtained from subject matter experts. In any development procedure that depends on judgments, it is important to tailor the instructions so that the resulting measure will reflect the task itself and only that. For example, in the construct weighting task, the low reliability of ratings provided by the noncommissioned officers may have been due to a confusion between rating numbers and ranking numbers.

Development of Measures

- Predictor development without job analyses is a chancy affair. It is true that there is an extensive literature on predictors of general abilities; it is also true that general predictors are much more administratively convenient in large-scale testing programs. However, the Job Performance Measurement project is devoting extensive resources to job specific criterion development. The use of general abilities as predictors has the effect of largely ignoring the job-specific information that was collected at fairly high cost. Military occupations that have a large number of incumbents or distinctive job requirements are likely to be better served by custom tailoring of predictors.
- Behaviorally anchored rating scales can include double-barreled questions to the extent that two or more concepts are included in the same scale. For example, in the construct weighting study, the rating scale for military bearing/physical fitness provides little guidance to the rater on how to evaluate physically fit soldiers with unkempt appearance.
- It must be kept firmly in mind that the research design endorsed by the Joint-Service Project is that hands-on performance tests should constitute the benchmark for criterion development and analyses. A major research goal of the Joint-Service Project is to evaluate the many different ways proposed to measure performance (e.g., hands-on performance tests, simulations, job knowledge tests, supervisory or peer ratings). It is likely that each way measures somewhat different aspects of performance and includes more or less

of the performance domain. Each way is also more or less costly. It is very important, therefore, to analyze each type of measure independently, in order to make informed decisions about substituting less intrusive and/or expensive measures for the benchmark hands-on measures. The data analysis from Project A lends support to this approach in that the hands-on measures were unrelated to the job and school knowledge tests (see the earlier section on the latent structure of performance).

- Analyses of job performance have tended to be dominated by a focus on relative differences among individuals rather than absolute performance capabilities of individuals. This is to be expected in a validation study, but there is no obvious way for such analyses to respond to questions such as: What performance scores indicate competent performance—or mediocre or expert performance—in a particular job? This latter question is of more pertinence to the larger goal of the Joint-Service Project: to help military policy makers set enlistment standards in the face of increasing budgetary pressures.

Data Reduction and Analyses

- A multiple trait/multiple method design is generally recognized as an exemplary research design. Experience has shown that method variance can be as large, and as important in its own right, as trait variance. The question has been one of what to do with method variance. The Army scientists uncovered two methods factors among their set of criterion measures, one for ratings and one for written tests. Partialling out these method variances from the corresponding criterion (trait) factors, to yield a clearer picture of these criterion factors, is innovative and potentially very useful. It is unfortunate that there were insufficient numbers of both hands-on and administrative measures to determine if these classes of measures too produced their own method variance.
- Large data sets lodged in powerful, high-speed computers provide an occasionally overwhelming temptation to investigators to put the machine through its paces rather than to select only those procedures that would be genuinely helpful in answering questions put to the data. It is always more appropriate to employ somewhat more thoughtful reflection and place somewhat less reliance on calculation. Such an approach is more likely to uncover meaningful relationships among variables as well as to be easier to explain to policy makers.

- With large data sets, investigators are tempted to reduce or cluster variables so as to minimize possible capitalization on chance in calculating statistical estimates of the relationships between classes of variables such as predictors and criteria. The appropriate posture is one of caution. At the very least, the rationale for such reductions needs to be explicitly stated, especially as the reduction may well proceed over several stages. All intermediate parameter estimates should be reported, as should be the choices made, and why. Nevertheless, multistage analyses of this type can sometimes impede discovery of more parsimonious or faithful representations of the latent structure of the variables in question.
- Constructs developed to summarize empirical research findings should reflect both the data and relevant theory. Positing constructs on the basis of military doctrine will not enhance understanding. The content of any military job is specified by policy decisions at the command level, that is, by doctrine. When measures are created and used to assess two kinds of activity (e.g., job-specific and common tasks), it should not be too surprising when these two activities emerge as inclusive factors. Such a conclusion may be more tautological than real.
- It is important to remember that capitalization on chance in data analyses can occur in a variety of ways and have unknown and profound effects. The traditional emphasis on cross-validation to deal with these problems bears reemphasis.
- Powerful data-handling procedures such as factor analysis each have their own idiosyncrasies, which may lead an investigator to accept a finding as fact when it is really an artifact. Several measures of one construct are required for a factor of that construct to emerge. Failure of a factor to appear may mean too few measures, not that the underlying construct fails to exist. Purely exploratory factor analysis is not necessarily appropriate without adequate representation of measures.
- In using LISREL, the investigator needs to remain aware that a causal relationship among intrapersonal variables is assumed. Predictors precede criteria, even if both types of measures are gathered at the same time. It is not known how robust LISREL is to the violation of its causal assumptions.
- Another concern with using LISREL is that successive reanalyses of the data, to make it fit a model refined and elaborated by these same data, would lead to an unwarranted and unknown capitalization on chance relationships in the data.

Part II

Proposals for Common Data Analysis

The December 1986 workshop provided an opportunity for all the Services and the committee to consider and discuss what sorts of analyses of the job performance measurement data might be most helpful and revealing. The Army, by describing its initial data analyses, provided a context for discussing the knotty problems facing the Joint-Service Project in interpreting and comparing the results of their research on job performance measurement.

The committee's proposals in this section, drawn from this discussion, address the major issue of coordinating the analyses across the Services, and thus across military occupational specialties and across testing methods, so that the results can be usefully interpreted. The results of the Joint-Service Project will be used for a variety of purposes by several communities, including the Services themselves, the U.S. Congress, the Department of Defense, and the general scientific community. They will all be best served if each Service reports its results in the same way so that all results are comparable.

The design of the Joint-Service Project itself requires that the Services adopt a common set of analyses. The comparison of various types of measures of job performance is one of the central project goals. All the Services have developed hands-on job sample measures, since these most closely approximate actual job performance. But most of the other measures being studied—paper-and-pencil tests of job knowledge, interview procedures, and simulations—are the province of one Service. To arrive at a conclusion about the relative

merits of each type of measure, the Services' research teams must adopt a set of common analytical strategies and identical methods of reporting results.

There are also sound policy reasons for a common data analysis plan. The Honorable David Armor, principal deputy assistant secretary (force management and personnel), in an address to workshop participants, called attention to the needs of the DoD for comparable data to use in supporting the Service requests to Congress for recruiting resources. He argued forcefully that DoD will be better able to substantiate quality requirements for Service personnel if the empirical findings derived from the job performance measurement research have the same meaning in all the Services.

A common approach to data analysis will also help the project to make a strong contribution to science. Measurement specialists will want to know all that has been learned about how to construct performance tests, as well as how *not* to construct them. They will want to know how hands-on tests stack up against alternatives such as walk-through performance tests, simulations, job knowledge tests, and supervisor's ratings. The profession will want to know the relative merit of measuring only job proficiency versus assessing total job performance—"can do" versus "does do." For the Joint-Service Project to make a coordinated contribution to scientific knowledge, results should be based on a common set of definitions and all data analyses should use the same methodology.

A common framework of data analysis is not advocated as a replacement, but rather as a supplement to Service-specific analyses. For some purposes it is appropriate for each Service to analyze data and report results in its own terms. For example, each Service's system of personnel selection and manpower management is unique, with its own data demands. In addition, the Army's Project A covers areas beyond the scope of the Joint-Service Project. But when the work *is* comparable, we urge that it also be *joint*, in the sense of using common definitions and common analyses, in order to make the maximal contribution to military policy and to science.

The text that follows sets forth a series of recommendations to the Job Performance Measurement Working Group for a core set of analyses of the job performance data collected by the Services over the last several years. The recommendations fall into two broad categories. The first and larger set deals with scientific evaluation of the performance data, including recommendations on scoring and scaling, handling missing data, reliability analysis, and exploring

score relationships. The concluding set of recommendations concerns analyses specifically designed to inform manpower policy. Definitions of the terms used in the recommendations follow this section.

The committee proposes that the results of these analyses be discussed at a workshop in fall 1988, capping a series of mutually beneficial exchanges between the committee and the participants in the Joint-Service Job Performance Measurement Project.

RECOMMENDATIONS: ANALYSES NECESSARY FOR THE SCIENTIFIC EVALUATION OF THE JOINT-SERVICE PROJECT

Performance Test Scoring

Recommendation 1: *Separate scores on criterion variables.* A separate total performance test score should be provided for each type of performance test. That is, a hands-on test score, a job knowledge test score, a walk-through performance test score, and so on, should be reported separately for each type of performance test or assessment used with a given job specialty.

A general principle of the committee's recommendations is that more detail is better than less. In the case of test scores, it might be tempting to combine scores to provide more stable, possibly more general information, but this may detract from our understanding what really happened. Moreover, one of the professed goals of the project is an evaluation of the relative merits of different testing formats. Comparisons can be made only if the results are reported separately for each format. There is good reason to expect quite different results from hands-on tests and job knowledge tests; one of the hypotheses underlying the Joint-Service Project is that hands-on tests give different and possibly more realistic information about performance than job knowledge tests.

Recommendation 2: *Job-specific test scores.* For the purpose of job classification studies, performance test scores should be based on job-specific tasks alone. If a test includes Service-wide tasks that all personnel are expected to know how to do no matter what their job assignment (e.g., cardiopulmonary resuscitation or firing a rifle), they should be

excluded from the job-specific score, unless they are specific to the job.

The Joint-Service Project has chosen to define job performance as job proficiency, as noted above. The main goal has been to determine what the incumbent "can do" as distinct from what he or she "does do." The main reason for this decision is to focus on information that would be useful in making classification decisions—which job the person should have, rather than whether the person should be accepted for military service. In line with this emphasis on differentiation, performance measures should provide job-specific scores. Including Service-wide items will tend to create more homogeneity in job performance than is desirable for job differentiation.

Recommendation 3: Common scale for performance tests. All performance test scores should be reported on a scale with the same range, with a lower limit of 0 and an arbitrary but standardized upper limit. We recommend an unfamiliar range, such as 0-150.

Putting all performance test scores on the same scale will avoid unwarranted assumptions that one test is more extensive than another because the scores have a wider range. This simple device of using the same scale for all jobs will emphasize that the scores themselves reflect how well an incumbent can do the job to which he or she has been assigned. The relative usefulness to the Service or the comparative quality requirements of different jobs are separate and later considerations.

Use of an unfamiliar scale is suggested simply to avoid comparison with other grade scales, as 0-100, 20-80, or the academic A, B, C, D, F. The scores should have a meaning intrinsic to their own definition, not because of their similarity to familiar scales.

Recommendation 4: Obtaining test scores from task scores. Performance test scores should be based on the unweighted or weighted combination of the task scores, item scores, or ratings. Any weights used should be reported and justified.

If the test has been assembled by a stratified random sampling of tasks, then the primary performance scale has been defined by the sampling weights; the selected tasks are considered representative of the job. Further weighting may be required to put all performance tests on the same 0-150 scale (such as if they are composed of different

numbers of tasks or items). Additional weighting is appropriate only to define alternative measures of performance based on a different rationale. For example, tasks could be selected on the basis of rated importance or criticality or both. Such differential weighting should be accompanied by detailed justification. All weighting should be explicit. More detailed discussion of task weighting is found in Wigdor and Green (1986:14-20).

> *Recommendation 5: Scaling task scores.* All task scores should be on the same scale, with a lower limit of 0, representing no skill at all, and an arbitrary upper limit, representing maximum performance. We suggest an upper limit of 10.

In order for task scores to be combined easily into a test score, as discussed above, with tasks receiving equal weight, the task scores themselves should be on the same scale. Then simply adding or averaging them will generally be appropriate. If task scores have different ranges, as for example, if the task score is the number of steps correctly done, and if some tasks have many more steps than others, then summing the task scores would mean that the tasks with more steps would be getting larger implicit weight. A task with 15 or 20 scoreable steps would overwhelm a task with only two or three scoreable steps.

One way to scale task scores is to score each step *go* or *no-go*; the task score is the proportion of steps scored *go*, multiplied by the established maximum score, e.g., 10. However, other scoring procedures are possible. A critical step approach might be used to determine pass/fail on some tasks, with failure at the critical step producing failure on the entire task. Alternatively, groups of steps might be weighted by importance. Steps may be separately scaled by group before combining. (For example, if the task were fixing a flat tire, three points might be given for removing the wheel, four points for fixing the flat tire, and three points for replacing the wheel, however many steps each part contains.) The critical feature is that all the task scores should have the same range. It is also important to distinguish between low or zero scores indicating low or minimal competence on the task and scores that are low because the data are missing. Strategies for obtaining task scores are discussed in detail in Wigdor and Green (1986:14-20).

Analysis of Performance Test Data

Recommendation 6: *Score distributions.* Frequency distributions of test scores, and the associated descriptive statistics (range, mean, median, standard deviation, quartiles, skewness, and kurtosis) are needed for all performance test scores and for the standard scores on each of the 10 subtests of the ASVAB, as well as for the Armed Forces Qualification Test and the Service composite, on the basis of which the recruits were selected for the job in question.

The nature of the relationship of the performance test scores to other variables is limited in part by the nature of the score distributions. Extremely skewed distributions, or very low variance, would lead to low relationships of the scores to other variables. The low relationships would not necessarily indicate an intrinsic lack of correlation but might be due only to the paucity of variance to relate to anything. Any unusual score distribution requires more detailed investigation to identify possible explanations.

Recommendation 7: *Missing data.* Values that are missing from the complete-data matrix should be multiply imputed using a two-imputation version of the impute algorithm used in Project A. The set of regression weights for estimating the missing data should be treated as a Bayesian system, each with a distribution of values. Two random draws from this multidimensional distribution should be made, yielding two sets of regression coefficients, from each of which a set of estimates is generated for the missing values. Both sets are used to fill out the data matrix, and each of the two resulting data matrices is subjected to the same data analyses. Finally, the results of these two analyses are averaged. Omitted responses that indicate failure to do a task should not be treated as missing data.

The Joint-Service Project requires extensive data collection with several hours of testing, sometimes over several days, for each incumbent. Illness, confusion, and logistical problems are inevitable. Performance data are so expensive to obtain, and the number of cases is so small, that every effort should be made to use whatever data are collected about each case. Complete-case analysis (Little and Rubin, 1987), which eliminates a case if any of its values is missing, is much too extreme, since most cases have some missing data.

The simplest way to keep the case (and to reach valid inferences) is to impute several values for each missing value. The data matrix can be analyzed using standard complete-data methods. Each set of imputed values results in a complete-data matrix, which is analyzed by standard complete-data methods, and the resulting analyses are combined as described in Rubin (1987). An existing algorithm for imputation used in Project A called IMPUTE is available from Wise and McLaughlin (1980); it should be modified to provide two or more imputations per missing value.

Other methods of treating missing data are possible, but in the interests of uniformity, we recommend using a common method, especially since current evidence indicates that differences would be slight.

> *Recommendation 8*: *Reliability.* An estimate of score reliability or of the variance of measurement errors should be provided for each type of test score. In addition, for each performance test that depends substantially on human scorers, an estimate of scorer reliability, or of the variance due to raters, should be provided. A generalizability analysis is preferred.

Any test is subject to errors of measurement. An index of reliability indicates the extent of such errors and is a standard requirement. The most common index of reliability is coefficient alpha, also known as KR-20, which reflects error variability due to test heterogeneity (i.e., task selection). When human scorers or raters must score the test, as with hands-on or walk-through performance tests, differences in human judgment are another source of score error and should also be evaluated. One method involves testing with parallel forms of the test using different scorers; one can then evaluate the errors attributable to scorers, item selection, and occasions, all together. A better method is a generalizability study (Shavelson, 1986), which would provide separate estimates of the size of each source of error.

Reliability estimates are a species of correlation, and as such are dependent on the variability of the measured group. Group variability must therefore be taken into account when comparing reliability coefficients for different methods that have been applied to different groups. Generalizability analysis avoids this problem by directly computing components of measurement error variance, in the metric of the test scores.

Recommendation 9: *Score relationships.* Correlations and covariances (uncorrected for selection and/or unreliability) are needed between every pair of variables formed from the set of all performance test scores, the standard scores on the 10 ASVAB subtests, the Armed Forces Qualification Test, and the Service composite for the job in question.

The extent to which the aptitude tests predict performance scores is the criterion-related validity of the predictors. The uncorrected covariances and correlations, together with the accompanying descriptive statistics of the score distributions (Recommendation 6), are fundamental to all further test analysis.

Recommendation 10: *Linearity of relationships.* For each pair of variables being correlated, the linearity of the relationship should be evaluated by comparing it with the nonlinear regression of each variable on the other.

If some variables are not linearly related to others, their correlations will not reflect the true nature of the relationship. For example, a performance test might require some minimal level of ability for success, but increasing amounts of ability past this point might not lead to increasing levels of performance.

Two ways of examining the linearity of relationship between two variables are through scatterplots and the descriptive correlation index *eta*. A scatterplot can show how values on a performance criterion increase, decrease, or stay the same for increasing values of an ability test predictor. The scatterplot can also show the relative dispersion of criterion scores for different predictor scores, or the degree of homoscedasticity. While scatterplots are very helpful in discovering and demonstrating the character of the relationship between two variables, their visual, nonverbal character limits their usefulness as descriptive or summary statements. The index *eta* is more helpful for these purposes.

Eta can be used to indicate the proportion of variance in one variable attributable to variation in the other variable. In other words, *eta* shows what changes in the predictor variable can be linked to changes in the criterion variable. Although the index *eta* is imprecise because adjacent score categories on the primary or predictor variable must be pooled when data are sparse, it is a useful index of relative size of effect. Homoscedasticity can be evaluated by examining the variance of the criterion scores within each predictor

score category. The difference between *eta* and the squared product-moment correlation indicates the proportion of correlated variance attributable to nonlinearity. An F-test can be used to determine if the size of this difference indicates a statistically significant departure from linearity, but a statistically significant effect that is nevertheless very small may be appropriately ignored.

> *Recommendation 11*: *Restriction of range.* A commonly accepted variance-covariance matrix of the predictor subtests should be used to correct the observed covariance matrices for range restriction. The covariances of the performance tests with the ASVAB tests should be corrected for range restriction using this matrix. It is recommended that the reference group for such a correction be the 1980 youth population (U.S. Department of Defense, 1982), omitting those with Armed Forces Qualification Test scores in the lowest 10 percent, who are ineligible for military service by congressional mandate.

Each job includes incumbents who are specially selected for that job and who choose that job. They are a more homogeneous group than the total applicant population, and their test scores will have less variance than those of the total group. Restricted test score ranges due to selection lead to lower validity correlations than would be found with the entire range of applicants. Furthermore, the size of the correlations depends to some extent on the degree of homogeneity of the predictor scores, which differs from one job to another. Thus the best and possibly the only way to compare validity correlations across jobs is to adjust them all to a common population. Because different methods of performance testing are being used on different jobs, cross-job comparisons are essential in order to compare the methods. Psychometricians call such adjustment to a common basis "correction for range restriction," because it is the range, which is to say the homogeneity of the group, that must somehow be equated.

Any such adjustment is an estimate, which depends on certain assumptions. Nevertheless, the need to compare testing methods requires a common basis, so the adjustment is necessary. In this project, the extent of the adjustment may be large; the larger the adjustment, the more the results depend on the correctness of the assumptions. A check of the assumptions, including the linearity of regression (Recommendation 10, above) and the homoscedasticity

of regression is necessary. Comparisons that use large adjustments should be made cautiously.

> *Recommendation 12*: *Effect of experience.* The Joint-Service Project should undertake an analysis relating job performance scores to experience, defined as months in service.

Interpretation of performance scores will be greatly enhanced by demonstrating their relationship to the incumbent's development on the job. Incumbents can be separated according to some objective measure of development. The one measure available to all the Services is time in service. We suggest separating the total group into at least three subgroups—up to 23 months, 24-35 months, and 36 or more months—and obtaining performance test score distributions for each group, so that the interaction of experience and performance variables can be displayed.

> *Recommendation 13*: *Subgroup analyses.* The amounts and types of relationships among scores, as detailed in Recommendations 9 and 10, should be considered separately for each experience subgroup if the data are sufficient. The amounts and types of relationships among scores for other subgroups of special interest, such as gender groups and various ethnic groups, should also be examined separately if the data are sufficient.

Detailed Analysis of Performance Tests

A detailed analysis of scores on tasks and steps on performance tests may provide insight into the interpretation of performance test scores. If interrelationships of task scores indicate that tasks appear to form clusters, subscores may be useful. Very low, or even negative, relationships between tasks would indicate that the job may be so heterogeneous that subspecialties are needed, or that de facto subspecialties exist. Performance test tasks and steps will require different analyses from the traditional item analysis appropriate for standardized test items. Job knowledge tests may be appropriately analyzed by the usual item analyses and item-total score correlations, but other types of tests will require more innovative approaches.

Detailed analysis of the performance test components will provide a useful record for later research and development efforts. Some of the results will certainly be task-specific or job-specific; other

results may generalize. Even the job-specific results will tend to indicate what sorts of considerations are crucial in designing performance tests.

> *Recommendation 14*: *Task-level item analysis.* For each task on the performance test, a distribution of task scores, including the descriptive statistics listed above (Recommendation 6) should be provided. The correlation of each task with the total test score, as well as the intertask intercorrelation matrix, should also be provided.

Since the task level scores range from 0 to 10, the usual item statistics (biserial correlations) are not relevant. Product moment correlations are appropriate. Task means and standard deviations will indicate the relative difficulty of the tested tasks. The intertask correlations will show if task clusters exist.

> *Recommendation 15*: *Task validity.* The correlations and covariances of the task scores with the ASVAB subtest scores should be provided, both unadjusted and adjusted for restriction of range.

Although such statistics may be unstable, it will be instructive to see if some tasks or some classes of tasks are better predicted from the aptitude and skill tests than others. Because of the need to compare across jobs, the range adjustment is appropriate. This may be pressing the range adjustment, but it seems necessary.

> *Recommendation 16*: *Step analysis.* For each task that is scored with steps (e.g., each hands-on task), a distribution of step scores should be provided, including the usual descriptive statistics (Recommendation 6), as well as the correlations of the step responses with the task scores.

The relationship of the step scores to the task score would be informative. It is important to note that for many types of tasks, the step scores are far from independent, so the step errors will not be independent. Many alternative analyses might be tried, including a Guttman scalogram analysis, but we have no specific guidance to offer.

RECOMMENDATIONS: ANALYSES USEFUL FOR INFORMING MANPOWER POLICY DECISIONS

Universal Job Scales

Recommendation 17: Common within-Service job scale. Each Service should devise a method of scaling different jobs on a single Service-wide scale. The common scale should represent in some form the relative standing of the various jobs.

Job allocation inevitably puts jobs in competition with each other. If an enlistee could become either a jet engine mechanic or a truck driver, the Service needs to know this, as well as how useful it would be to assign him or her to each job.

In addition, if the applicant could become either a mediocre jet engine mechanic or a top-notch truck driver, the Service needs to know this as well, and how to attach priorities to each. Furthermore, the enlistee's preferences should somehow be taken into consideration.

There is at present no accepted way of scaling jobs, either to some expected criterion levels of performance within one job or to criterion levels across jobs, so we advocate separate work by each Service in coordination with its allocation system.

Recommendation 18: Universal job scale: across Services. Eventually a single scale should be selected to serve as the framework for specifying manpower quality requirements, as well as justifying requests for recruiting resources.

Definitions

Job A military job; a *military occupational specialty* (MOS) in the Army and Marines, a *rating* in the Navy, or an *Air Force Specialty* (AFS) in the Air Force.

Performance Test Components

Task A job consists of a large number of tasks, each of which is a circumscribed, well-defined unit of activity. (A job may be viewed in many ways; for purposes of test development the Joint-Service Project has chosen to view a job as a collection of tasks.)

Step A task used as a test item generally involves a series of successive actions, or *steps*. A step is an identifiable unit, usually a scoreable unit. In actual job performance, tasks may not always have the same steps, the same number of steps, or the same sequence of steps. Here we discuss steps only in the tasks as tested.

Item In this context, *item* means a scoreable unit—it could be a task, a step, or a question on a written test.

Criterion Measures

Performance Test A test consisting of several different tasks, selected from the total domain of job tasks, on which each examinee is evaluated. The selected tasks are intended to be representative of the job content.

Hands-on Performance Test Each tested task in a hands-on test is evaluated by having the examinee actually do a particular

instance of the task or a very close replica of the real task. A hands-on performance test is a specific type of performance test.

Walk-through Performance Test A walk-through performance test combines hands-on and interview procedures. Some tasks are actually performed; for others, examinees merely describe how they would do the task.

Simulation Performance Test In a simulation performance test, each tested task is a detailed simulation of a real task. For example, an air traffic controller would most naturally be tested on simulations of actual air situations, rather than the real thing. Some simulations have relatively low fidelity, including many paper-and-pencil analogs; high-fidelity simulations often involve special equipment.

Comment: Occasionally some items on a hands-on performance test are simulations or have a walk-through format; e.g., the fireman's carry is performed with a dummy; for applying field dressings, a confederate pretends that his hand has been amputated. In general, the nature of the test is defined by the nature of the preponderance of its tasks.

Job Knowledge Test A test of job knowledge is a series of items, usually independent, usually multiple choice, concerning specific knowledge of some part of a job. An item on a job knowledge test may correspond to a certain task or to a specific step in a task, but it may also relate to a more global aspect of the job.

Selector or Predictor Measures

Aptitude Test Aptitude tests refer to the ASVAB and its various subscores, as well as aptitude tests currently being explored as additions to the ASVAB. They are intended to be measures of the underlying abilities required on the job and are used to predict levels of individual job performance.

References

Campbell, John P.

1986 Validation Analysis for New Predictors. RS-WP-86-09. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Va.

Committee on the Performance of Military Personnel

1984 *Job Performance Measurement in the Military: Report of a Workshop.* Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, D.C.: National Academy Press.

Little, Roderick J.A., and Donald B. Rubin

1987 *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

McHenry, Jeffrey J., Lauress L. Wise, John P. Campbell, and Lawrence M. Hanser

1986 A Latent Structure Model of Job Performance Factors: Appendix. RS-WP-86-10. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Va.

Personnel Decisions Research Institute

1985 Task 2 Scientific Advisory Group: Cognitive Domain. Briefing prepared by Marvin D. Dunnette. Personnel Decisions Research Institute, Minneapolis, Minn., March 20-21.

Rubin, Donald B.

1987 *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Rumsey, Michael G., William C. Osborn, and Patrick Ford

1985 Comparing Work Sample and Job Knowledge Measures. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Va. Paper presented at the annual meeting of the American Psychological Association, Los Angeles, Calif., August 23-27.

Sadacca, Robert, Maria Veronica de Vera, and Ani S. DiFazio

1986 Weighting Performance Constructs in Composite Measures of Job Performance. Paper presented at the convention of the American Psychological Association, Washington, D.C., August 22-25.

Schmitz, Edward J.
1987 Project B: Improving Personnel Performance Through Assignment Policy. Paper presented at the Workshop on Linkage Issues, Committee on the Performance of Military Personnel, Santa Fe, N.M., June 22-26.

Shavelson, Richard J.
1986 Generalizability of Military Performance Measurements: I. Individual Performance. Paper prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education. National Research Council, Washington, D.C.

U.S. Department of Defense
1982 *Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery.* Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics). Washington, D.C.: U.S. Department of Defense.

Wigdor, Alexandra K., and Bert F. Green, Jr., eds.
1986 *Assessing the Performance of Enlisted Personnel: Evaluation of a Joint-Service Research Project.* Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, D.C.: National Academy Press.

Wing, Hilda, Leaetta M. Hough, and Norman G. Peterson
1987 Predictive Validity of Noncognitive Measures for Army Classification and Attrition. Paper presented at the 2nd annual conference of the Society for Industrial and Organizational Psychology, Inc., Atlanta, Ga., April.

Wing, Hilda, Norman G. Peterson, and R. Gene Hoffman
1984 Expert Judgments of Predictor-Criterion Validity Coefficients. Symposium presented at the convention of the American Psychological Association, Toronto, Ontario, August.

Wise, Lauress L., John P. Campbell, Jeffrey J. McHenry, and Lawrence M. Hanser
1986 A Latent Structure Model of Job Performance Factors. Paper presented at the convention of the American Psychological Association, Washington, D.C., August 22-25.

Wise, Lauress L., Jeffrey J. McHenry, and Winnie Y. Young
1986 Project A Concurrent Validation: Treatment of Missing Data. RS-WP-86-08. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Va.

Wise, Lauress L., and D. McLaughlin
1980 *Guidebook for the Imputation of Missing Data.* Palo Alto, Calif.: American Institutes for Research.

Attendance

JANE M. ARABIAN, Research Psychologist, Selection and Classification Technical Area, U.S. Army Research Institute for the Behavioral and Social Sciences

DAVID J. ARMOR, Principal Deputy Assistant Secretary of Defense (Force Management and Personnel)

JERALD G. BACHMAN, Survey Research Center, University of Michigan*

HERBERT BAKER, Navy Personnel Research and Development Center

LT. COL. RODGER D. BALLENTINE, Chief, Skills Development Branch, Training Systems Division, Air Force Human Resources Laboratory

V. JON BENTZ, Director (Ret.), Psychological Research and Services, Sears, Roebuck & Company, Elmhurst, Ill.*

LLOYD BOND, Learning Research and Development Center, University of Pittsburgh*

JOHN P. CAMPBELL, Principal Scientist, Army Project A; Professor of Psychology, University of Minnesota, Twin Cities

N. KENT EATON, Director, Manpower and Personnel Research Laboratory, U.S. Army Research Institute for the Behavioral and Social Sciences

ROBERT L. FREY, Personnel Research Psychologist, Headquarters, U.S. Coast Guard

DIANE L. GOLDMAN, National Research Council*

*Committee on the Performance of Military Personnel

BERT F. GREEN, JR., Department of Psychology, Johns Hopkins University*
LAWRENCE M. HANSER, Chief, Selection and Classification Technical Area, U.S. Army Research Institute for the Behavioral and Social Sciences
LT. COL. DICKIE HARRIS, Program Manager, Joint-Service Job Performance Measurement Project
JAMES H. HARRIS, Deputy Director, Army Project A; Human Resources Research Organization
JERRY W. HEDGE, Chief, Performance Measurement Section, Air Force Human Resources Laboratory
COL. WM. DARRYL HENDERSON, Commander, U.S. Army Research Institute for the Behavioral and Social Sciences
MAJ. LAWRENCE JURICA, Headquarters, U.S. Marine Corps
LEONARD P. KROEKER, Personnel Research Psychologist, Navy Personnel Research and Development Center
GERALD J. LAABS, Acting Division Head, Personnel Systems Research Department, Navy Personnel Research and Development Center
ROBERT L. LINN, School of Education, University of Colorado*
MILTON H. MAIER, Center for Naval Analyses
PAUL W. MAYBERRY, Center for Naval Analyses
JEFFREY J. McHENRY, Washington Research Center, American Institutes for Research
MAJ. GEN. WILLIAM G. O'LEKSY, Director of Military Personnel Management, U.S. Army
JOHN W. ROBERTS, U.S. Air Force (Ret.), San Antonio, Texas*
HENDRICK W. RUCK, Technical Advisor, Training Systems Division, Air Force Human Resources Laboratory
ROBERT SADACCA, Technical Planning and Research Design, Army Project A; Human Resources Research Organization
MADY W. SEGAL, Department of Sociology, University of Maryland*
W.S. SELLMAN, Director for Accession Policy, Office of the Assistant Secretary of Defense (Military Manpower and Personnel Policy)
RICHARD J. SHAVELSON, Dean of Graduate Education, University of California, Santa Barbara*

*Committee on the Performance of Military Personnel

LT. COL. WILLIAM J. STRICKLAND, Chief, Accession Testing Policy, Directorate of Personnel Plans, Headquarters, U.S. Air Force/DPXOA

MARY L. TENOPYR, Army Scientific Advisory Group; AT&T Corporate Headquarters, New York

LONNIE D. VALENTINE, JR., Chief, Force Acquisition Branch, Air Force Human Resources Laboratory

H.P. VAN COTT, Study Director, Committee on Human Factors, National Research Council*

ALEXANDRA K. WIGDOR, National Research Council*

LAURESS L. WISE, Director for Analysis, Project A; American Institutes for Research

*Committee on the Performance of Military Personnel

National Academy Press

The National Academy Press was created by the National Academy of Sciences to publish the reports issued by the Academy and by the National Academy of Engineering, the Institute of Medicine, and the National Research Council, all operating under the charter granted to the National Academy of Sciences by the Congress of the United States.

BEST COPY AVAILABLE